GATEKEEPER ARENA: DIAGNOSING PERCEPTION AND MEMORY FOR MULTIMODAL POLICY DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Policy decisions can fail when relevant state is absent from the current encounter. We present **Gatekeeper Arena**, an offline testbed with 100 fictional entrants, 222 Unreal Engine clips, a disclosed nine-role policy, and a three-stage lifecycle: profile registration, unscored familiarization, and frozen-memory evaluation. Models return a fixed record; event-level balanced accuracy becomes the 0–100 Arena Score, accompanied by decision coverage. We evaluate the current-evidence setting with three open-weight VLMs across all five text/visual conditions, and ten hosted families join them on a shared text-plus-eight-frame leaderboard (13 models; 2,925 scheduled decisions). All cluster near chance (49.78–54.01) and are strongly deny-biased. FlyWire-v783 recurrence retains delayed synthetic input, yet its Arena Score is 45.25 versus 50.53 for a degree-preserving rewire. Additional frames and generic recurrence therefore do not supply the profile and encounter state required by a longitudinal gatekeeper.

1 INTRODUCTION

Multimodal benchmarks usually ask questions about visible content (Yue et al., 2024; Li et al., 2023; Fu et al., 2025). A policy system must instead act even when the target is not identifiable from supplied evidence. Fluency is insufficient: false admits and rejects differ, unresolved outputs require coverage accounting, and repeated encounters may require hidden state.

At a simulated residential gate, an offline model predicts the action for a fictional entrant. Authored role and prior encounter state are simulator variables rather than visual annotations. Moreover, 81 of 100 entrants have allow-policy roles, so always-allow reaches 81% accuracy while recognizing no denied case. Raw accuracy, extra frames, or conservative denial can therefore mislead.

Gatekeeper Arena separates the lifecycle from two diagnostic axes (Figure 1). Before deployment, the agent receives labelled resident profile cards; during one unscored familiarization day it may append encounter notes; the resulting memory is then frozen for formal evaluation. *Track P* varies present text and frames with empty longitudinal context. *Track C* measures delayed-input capacity under recurrent-topology changes and tests transfer back to the same gate-policy endpoint. This distinguishes visible evidence, longitudinal state, and generic recurrence.

All thirteen open/hosted models cluster near chance on shared T+V-8 events and recover few allows. FlyWire retains delayed input while its original topology trails the rewired control on the gate task. Together these results localize the bottleneck to task-relevant longitudinal state.

Identity-disjoint folds, explicit decision coverage, frozen prompts, and one-shot test access structure the comparison.

Our contributions are:

- an offline 100-entrant, 222-clip policy benchmark that pairs controlled evidence conditions with event-level Arena Score and directional errors across open and hosted models;
- an explicit three-stage protocol that separates registered profile state, writable familiarization memory, and read-only formal evaluation; and
- a capacity–transfer counterexample showing that topology-sensitive synthetic memory need not improve the downstream policy decision.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

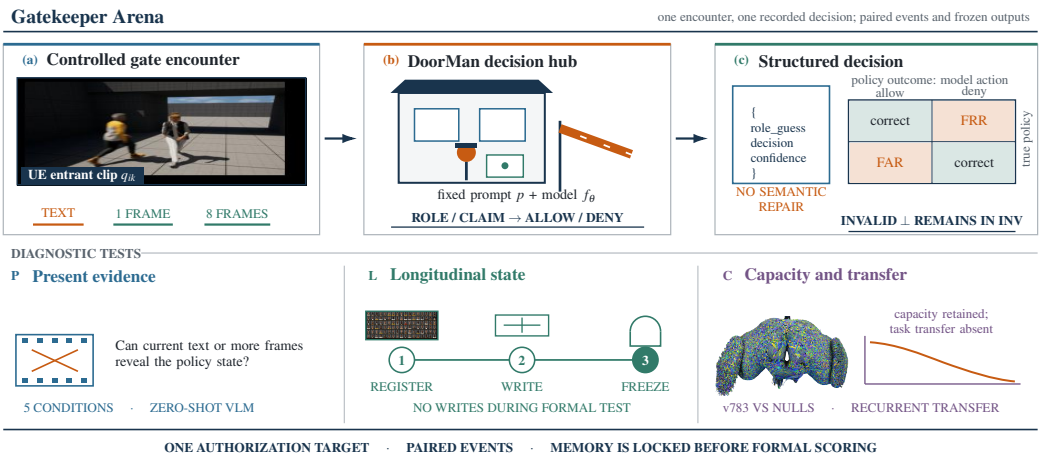

Figure 1: Gatekeeper Arena as a single visual story. A synthetic entrant approaches the DoorMan decision hub, which emits one recorded outcome: correct, false admit (FAR), false reject (FRR), or invalid $\perp$; invalid responses remain in coverage/INV rather than being assigned an action. The lower row isolates present evidence, longitudinal memory, and recurrent capacity. FlyWire: Tyler Sloan, Princeton University (Dorkenwald et al., 2024).

2 Related Work

From multimodal answers to decisions. General image/video suites measure perception and reasoning (Yue et al., 2024; Liu et al., 2024b; Yu et al., 2024; Fu et al., 2025; Li et al., 2024; Mangalam et al., 2023); social benchmarks target understanding or action (Zadeh et al., 2019; Jin et al., 2024; Lin et al., 2025; Mathur et al., 2025; Zhou et al., 2024; Xu et al., 2026; Zhang et al., 2025). Gatekeeper Arena occupies a complementary point (Table 1): it is offline and intentionally narrow, but makes the policy, repeated-identity intervention, directional errors, and failure accounting part of the task definition rather than post-hoc analysis.

Table 1: Closest benchmark families, compared by explicit evaluation axes rather than perceived quality. ✓: primary design feature; △: context is present but not an identity-controlled intervention; –: not a primary axis in the cited benchmark.

Benchmark family	Primary input	Endpoint	Repeated identity	Asymmetric errors	Invalids retained
MMMU / MMBench / MM-Vet	image + text	answer	–	–	–
Video-MME / MVBench / EgoSchema	video + text	answer	–	–	–
Social-IQ / MMToM-QA	video/dialogue	social answer	–	–	–
LoCoMo / LongMemEval	dialogue history	recall answer	△	–	–
SOTOPIA / visual-action suites	context / vision	action	△	△	–
Gatekeeper Arena	matched text/video	policy action	✓	✓	✓

Grounding and longitudinal memory. VQA studies expose answer priors and modality failures (Goyal et al., 2017; Agrawal et al., 2018; Gat et al., 2021); annotation artifacts also arise in language-only inference data (Gururangan et al., 2018). We instead hold the encounter fixed while withholding text or frames. Long-horizon memory benchmarks mostly study conversational recall (Maharana et al., 2024; Wu et al., 2025; Cheng et al., 2026). Person re-identification studies cross-view identity retrieval (Zheng et al., 2016), while open-set recognition and selective classification formalize unknown inputs and abstention (Bendale & Boult, 2016; Geifman & El-Yaniv, 2017). Our question is operational: can an agent use labelled profile state and unscored encounters without leaking formal-test labels? We therefore make the write window, snapshot boundary, read-only test phase, and identity-disjoint event folds part of the benchmark contract. The empirical matrix instantiates its current-evidence, empty-context condition.

Recurrent and connectome-constrained computation. Reservoir computing measures how a fixed dynamical system exposes past input to a simple readout (Lukoševičius & Jaeger, 2009; Dambre et al., 2012). FlyWire adds a whole adult *Drosophila* wiring diagram and neuron annotations (Dorkenwald et al., 2024; Schlegel et al., 2024); recent work uses anatomy to constrain sensorimotor and visual models (Shiu et al., 2024; Lappalainen et al., 2024; Wang-Chen et al., 2024). We use the connectome as a fixed recurrent control: first test delayed-input capacity against matched graph nulls, then test whether that capacity transfers to gate decisions.

Reliable evaluation. Trustworthiness suites broaden the failure surface of multimodal models (Zhang et al., 2024; Liu et al., 2024a), while HELM, datasheets, and data cards emphasize multi-metric reporting and explicit documentation (Liang et al., 2023; Gebru et al., 2021; Pushkarna et al., 2022). Our narrower contribution retains unresolved events in coverage accounting and separates Arena Score, directional errors, and platform availability.

3 Gatekeeper Arena

3.1 Task Definition

An Arena instance is the tuple $\mathcal{A} = (\mathcal{I}, \mathcal{Q}, \mathcal{R}, \pi, \mathcal{P}, \mathcal{F}, \mathcal{E})$. $\mathcal{I}$ is a set of fictional entrant identities, $\mathcal{Q}_i$ the rendered clips belonging to identity i, and $\mathcal{R}$ nine authored roles. The disclosed policy π maps structured event state—including role and, when present, identity-claim validity—to allow (0) or deny (1). $\mathcal{P}$ is the pre-operational resident-profile registry, $\mathcal{F}$ the unscored familiarization stream, and $\mathcal{E}$ the scored stream after memory is frozen. An encounter is $e_{ik} = (i, q_{ik}, u_{ik}, r_i, h_{ik}, t_{ik})$, where $q_{ik} \in \mathcal{Q}_i$ is a clip, u_{ik} is its optional utterance, r_i is authored metadata, h_{ik} is optional claim state, and t_{ik} is the event time. The target is $y_{ik} = \pi(r_i, h_{ik})$. An evidence condition c exposes a deterministic view $x_{ik}^{(c)}$—text, one frame, eight frames, or either visual view with text—while holding identity, clip, target, prompt, and decoding fixed.

The model emits one three-field record $o_{ik} = (\tilde{r}_{ik}, \tilde{y}_{ik}, c_{ik})$: predicted role, allow/deny decision, and self-reported confidence in that decision. A strict parser maps the decision to $\hat{y}_{ik} \in \{0, 1, \perp\}$, where $\perp$ denotes a missing, malformed, or illegal decision. It may strip one Markdown fence but performs no semantic repair. Policy correctness, role correctness, confidence coverage, and record validity therefore remain separate observables. Figure 1 visualizes this contract and the three diagnostic interventions.

The authored role is latent simulator metadata. The lifecycle supplies that state through labelled profile cards, familiarization observations, and a read-only evaluation snapshot. Track P sets the snapshot to empty, isolating what current text and pixels contribute. Always-allow and always-deny are imbalance controls; an oracle given r_i defines the task ceiling.

3.2 Population, Corpus, and Policy

We use *gate entrant* for any depicted person; *owner* denotes only the exact authored owner role. The corpus contains 100 fictional profiles and 222 unique Unreal Engine approach clips. The inventory records each clip's identity, duration, frame rate, and resolution. Sixty-three identities have at least two distinct clips, enabling query/support separation.

Each profile has authored attributes, a short fictional history, and one role. Figure 2 reports composition computed from source records. The policy allows six roles (81 identities) and denies three (19). Thus always-allow obtains 81% raw accuracy but zero deny recall, whereas both constant policies obtain Arena Score 50. The open-set Track P schedule additionally pairs clips with matched or mismatched claims; its event-class prevalence is therefore distinct from the 81/19 profile composition. Arena Score balances either binary view.

Scores measure conformance to this fictional policy. Its deny set includes an exclusionary "vagrant" label, motivating the fairness discussion in Section 6.

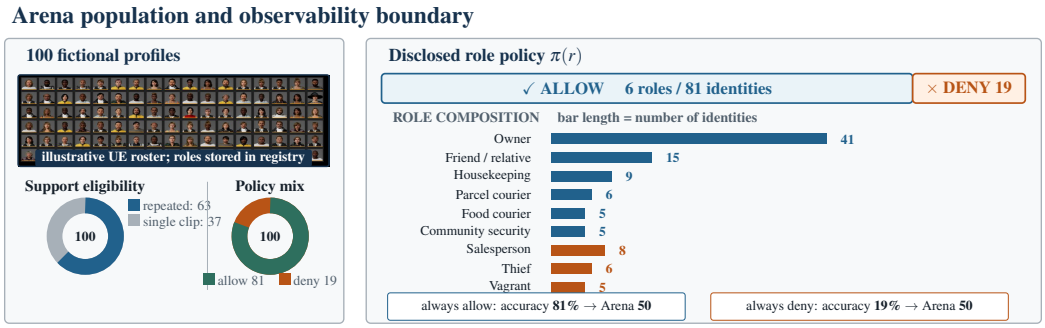

Figure 2: Verified corpus and policy composition. Donuts show repeated-clip eligibility and the 81/19 policy mix; the role panel accounts for all 100 fictional identities. Constant policies expose raw accuracy's failure; role labels come from the simulator registry.

3.3 Three-Stage Longitudinal Protocol

The benchmark follows the deployment cycle in Figure 3. **Phase 1 (registration)** presents the resident registry before any scored encounter; a card contains the fictional identity label, apartment metadata, and a synthetic portrait reference. **Phase 2 (familiarization)** replays one logical day of unscored encounters. Memory writes are append-only, individually audited, and limited to events scheduled before the freeze boundary. **Phase 3 (formal evaluation)** freezes the resulting snapshot and replays seven logical days of scored events; evaluation events cannot mutate identity or policy state. "Day" denotes a schedule partition because the available assets are inbound clips, not continuous UE trajectories.

The protocol pairs an empty-context view with a frozen-snapshot view on the same ordered events. Track P instantiates the empty-context view; Track C tests whether recurrent state alone transfers to the same decision metric.

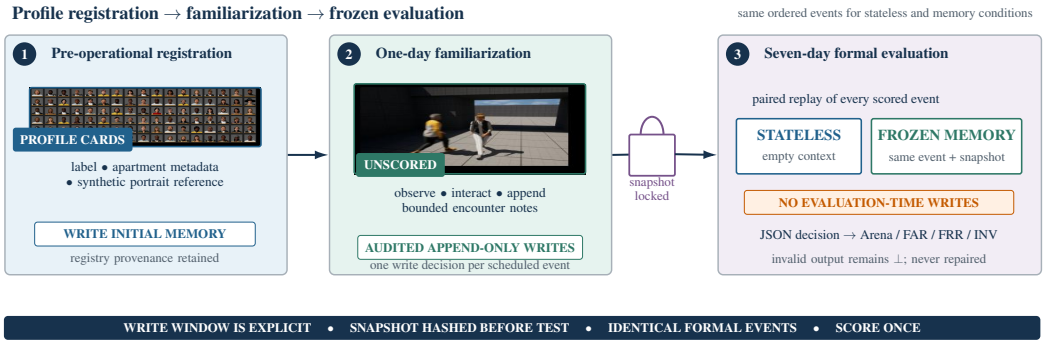

Figure 3: Longitudinal Arena contract. Resident profile cards are registered before deployment, one unscored familiarization day may append audited observations, and the resulting snapshot is locked before seven logical evaluation days. Stateless and frozen-memory conditions share each scored event; only the memory context differs.

3.4 Track P: Zero-Shot Policy Decisions

The protocol gives each model one Chinese instruction naming the closed role set, policy, and three-field output contract. Outputs are validated field by field. An otherwise valid decision is still scored when another field fails; malformed or non-object output yields $\hat{y} = \perp$. For hosted execution, a valid single-worker recheck replaces a failed first response when available. Remaining $\perp$ events contribute to class-specific and overall decision coverage; Arena/FAR/FRR use resolved decisions.

Five fixed identity-disjoint folds cross the five evidence conditions for the open-weight ablation. Frame indices are selected uniformly and deterministically. Text-only omits video; vision-only

omits the utterance. The resulting contrasts are conditional evidence ablations. The five folds contain 225 scheduled encounters per condition. Thus each fully evaluated model produces $225 \times 5 = 1{,}125$ decisions over 25 model–fold–condition cells, and the three full-matrix models contribute 3,375 decisions. Multiple clips of one identity remain one statistical cluster.

Table 2: Frozen evidence conditions. Every condition reuses the same ordered identity/clip schedule; only the supplied evidence changes.

Condition	Supplied evidence	Diagnostic question
Text	Utterance	Is language alone sufficient?
V-1	One sampled frame	What is visible instantaneously?
V-8	Eight ordered frames	Do more visual observations help?
T+V-1	Utterance + one frame	Does minimal fusion help?
T+V-8	Utterance + eight frames	Does the richest present evidence help?

Because utterances may cue occupation or purpose, text-only performance is an explicit condition rather than neutral context; models see neither profile cards nor the frozen longitudinal snapshot.

The representative instruction is: "inspect the eight ordered frames, apply the disclosed nine-role policy, and return only `{role_guess, decision, confidence}`, with no explanation or Markdown." The exact Chinese prompt and parser domains are released; this same contract is held fixed across models.

The cross-access leaderboard freezes T+V-8 and the same three-field JSON record (role guess, decision, confidence). It uses five open-set folds whose binary authorization labels include hidden identity-claim state; consequently the same visual clip can occur under different targets when the claim changes. The empty-context condition withholds that claim, creating a controlled missing-state diagnostic. The three open-weight models and ten hosted families receive the same ordered events, source clips, textual context, prompt, and targets. The local stack samples uniformly over the decoded span, whereas the hosted stack pre-extracts eight deterministic interior frames (5–95%); access is shown in the table. This yields 225 scheduled events per model and 2,925 total. INV records unresolved events, and uncertainty clusters by fold and visual identity. Open checkpoints are pinned to public revisions; hosted rows retain the requested and returned family aliases.

3.5 Track C: Connectome Capacity and Transfer

FlyWire v783 is instantiated as a signed recurrent operator driven by synthetic sequences. For Rademacher input $u_t \in \{-1, +1\}$, its signed, postsynaptic-row-normalized adjacency evolves as

$$x_{t+1} = \tanh(0.9Wx_t + 0.1bu_t). \tag{1}$$

The input vector b addresses 256 deterministic afferents with balanced signs; only 1,481 efferent coordinates are exposed to a frozen ridge readout. Each 4,096-step sequence uses 256 washout and 2,048/768/1,024 train/validation/test states. Memory capacity is $\mathrm{MC} = \sum_{d=1}^{16} \mathrm{corr}^2(u_{t-d}, \widehat{u}_{t-d})$, following the linear-memory-capacity construction (Dambre et al., 2012).

Four graph conditions isolate structure: original FlyWire (C0), a directed degree-preserving rewire retaining source-wise signed weights (C1), a source-wise weight shuffle on fixed edges (C2), and a degree-stratified afferent/efferent boundary permutation (C3). All nulls reuse the same five sequence seeds. Appendix A.2 fixes the complete indexing and readout contract. We then freeze per-event reservoir features and compare C0, C1, a learned GRU, and a memoryless MLP over five seeds and 1,600 test events per cell. The two stages deliberately distinguish "can retain input" from "improves the policy task."

4 Evaluation

4.1 Unit of Analysis and Primary Estimand

The scheduled encounter is the accounting unit and visual identity is the dependence cluster. Let $\mathcal{S}_D$ and $\mathcal{S}_A$ be all scheduled deny and allow events, and let $\mathcal{E}_D \subseteq \mathcal{S}_D$ and $\mathcal{E}_A \subseteq \mathcal{S}_A$ contain events with

a resolved legal decision after the fixed response-resolution rule. Deny and allow recall are

$$R_D = \frac{1}{|\mathcal{E}_D|} \sum_{(i,k)\in\mathcal{E}_D} \mathbb{1}[\hat{y}_{ik} = 1], \quad R_A = \frac{1}{|\mathcal{E}_A|} \sum_{(i,k)\in\mathcal{E}_A} \mathbb{1}[\hat{y}_{ik} = 0], \quad \mathrm{BAcc} = \tfrac{1}{2}(R_D + R_A), \tag{2}$$

where $\hat{y}_{ik} \in \{0, 1\}$ for resolved events. Balancing the two event classes prevents the 81/19 role composition from dominating the score. Bootstrap resampling keeps every visual-identity cluster intact. BAcc therefore summarizes equal-class decision capability.

For a readable same-protocol leaderboard, we define

$$\mathrm{ArenaScore} = 50(R_D + R_A) = 100\,\mathrm{BAcc}. \tag{3}$$

The score ranges from 0 to 100 and gives equal weight to correct deny and correct allow decisions. It is a direct rescaling of the primary endpoint. Rankings compare models under an identical task and evidence condition, with intervals shown for every row. All policy-decision tables and plots use Arena points on 0–100; FAR, FRR, and unresolved rate (INV) use percentages. Connectome memory capacity is reported on its native scale.

Metric contract. ↑ is better and ↓ is better; Arena Score is BAcc on a 0–100 scale.

Metric	Question	Aggregation	Invalid handling
BAcc ↑ (primary)	Is the policy action correct?	Resolved event, then policy class	⊥ excluded; coverage reported
Arena Score ↑	Same endpoint on a 0–100 scale	100×BAcc	Same resolved set
FAR / FRR ↓	Which direction fails?	Resolved rate within true class	⊥ reported separately
$\mathrm{INV}_{A/D}$ ↓	Is the decision resolved?	Scheduled rate within class	Is the endpoint
Role macro-F1 ↑	Is the authored role recovered?	Macro over nine roles	Failed role is incorrect

4.2 Directional Errors and Decision Coverage

Malformed records map to ⊥. FAR and FRR are computed among resolved events, while unresolved rates retain the scheduled class denominators:

$$\mathrm{FAR} = \frac{1}{|\mathcal{E}_D|} \sum_{(i,k)\in\mathcal{E}_D} \mathbb{1}[\hat{y}_{ik} = 0], \qquad \mathrm{INV}_D = 1 - \frac{|\mathcal{E}_D|}{|\mathcal{S}_D|}, \tag{4}$$

$$\mathrm{FRR} = \frac{1}{|\mathcal{E}_A|} \sum_{(i,k)\in\mathcal{E}_A} \mathbb{1}[\hat{y}_{ik} = 1], \qquad \mathrm{INV}_A = 1 - \frac{|\mathcal{E}_A|}{|\mathcal{S}_A|}. \tag{5}$$

Consequently $R_D + \mathrm{FAR} = 1$ and $R_A + \mathrm{FRR} = 1$ on resolved events. $\mathrm{INV}_{A/D}$ exposes class-specific missingness.

Track P confidence accompanies the chosen action and is summarized through field coverage. The compact contract above fixes the denominator and invalid handling for every reported metric.

4.3 Uncertainty and Multiplicity

For the Track P evidence ablation, the fold is the uncertainty unit because identities are disjoint across folds. We report fold ranges and exact paired permutation tests for five predeclared evidence contrasts with Holm correction. Displayed point estimates pool confusion counts over the five folds before computing event-level Arena Score; this differs from averaging five fold-level BAcc values. The unified T+V-8 leaderboard reports the same pooled point estimate with a hierarchical bootstrap over folds and then visual-identity clusters for every open and hosted row. For Track C, memory capacity is the sum of delayed-input reconstruction scores. Graph nulls share sequence seeds and task-level arms share five seeds, so effects are paired.

The same analysis code computes all reported metrics, intervals, and figures. Appendix A specifies the scoring rules and experimental details needed to reproduce them.

Table 3: Main results. **A:** open-weight evidence ablation on five frozen folds (225 decisions per condition). **B:** response-resolved open/hosted T+V-8 leaderboard on shared events, clips, targets, and output fields, followed by a FlyWire transfer reference under its own frozen protocol. Arena Score is $100\times$BAcc; FAR, FRR, and INV are percentages. ↑ is better; ↓ is better. Bold marks the highest value within each block.

A. Open-weight evidence ablation (Arena Score ↑)

	Arena Score (50 = chance)					T+V-8 error rates (%)		
Model	Text	V-1	V-8	T+V-1	T+V-8	FAR	FRR	INV
Qwen2.5-VL-3B[a]	50.1	50.0	49.3	50.4	50.3	14.0	**85.4**	0.0
Qwen2.5-VL-7B[a]	**50.8**	**50.4**	**50.6**	**51.4**	**50.3**	**4.2**	95.1	0.0
Qwen3-VL-8B[b]	**50.8**	50.0	49.2	49.1	50.1	13.3	86.6	0.0

B. Response-resolved open-set T+V-8 leaderboard

Rank	Model family	Access	Arena ↑	95% CI	FAR ↓	FRR ↓	INV ↓
1	Qwen3.8 Max[l]	H	**54.01**	[49.7,58.8]	10.5	81.5	0.4
2	GPT-5.6 Luna[e]	H	52.41	[47.3,58.2]	14.7	80.5	**0.0**
3	Grok 4.6[j,†]	H	51.88	[45.4,57.7]	23.4	**72.8**	1.3
4	Claude Opus 5[h]	H	51.86	[46.8,56.4]	14.8	81.5	0.9
5	GPT-5.6 Terra[e]	H	51.19	[46.5,56.4]	14.7	82.9	**0.0**
6	Claude Fable 5[g]	H	50.95	[46.3,55.2]	15.6	82.5	1.8
7	GPT-5.6 Sol[e]	H	50.79	[45.8,56.4]	15.5	82.9	0.4
8	GPT-6 Astra[f]	H	50.37	[46.0,54.5]	14.1	85.2	0.9
9	Qwen2.5-VL-7B[a]	O	50.34	[46.8,52.6]	**4.2**	95.1	**0.0**
10	Qwen2.5-VL-3B[a]	O	50.32	[46.2,55.0]	14.0	85.4	**0.0**
11	Gemini 3.1 Pro Preview[i,†]	H	50.15	[46.3,53.9]	11.9	87.8	**0.0**
12	Qwen3-VL-8B[b]	O	50.06	[45.8,54.3]	13.3	86.6	**0.0**
13	Kimi-K3[k,†]	H	49.78	[44.4,55.5]	17.7	82.7	1.3
Arena-standard connectome transfer reference (1,600-event protocol)							
–	FlyWire v783 C0 + MLP[m]	C	45.25	[42.2,48.0]	74.7	34.8	0.0

O: public open-weight checkpoint; H: hosted family requested through the unified platform. INV is the residual unresolved percentage after available single-worker rechecks; those events are excluded from Arena/FAR/FRR. † Recheck details appear in Appendix C. Model sources: [a](Bai et al., 2025b); [b](Bai et al., 2025a); [e](OpenAI, 2026a); [f](OpenAI, 2026b); [g](Anthropic, 2026a); [h](Anthropic, 2026b); [i](Google DeepMind, 2026); [j](xAI, 2026); [k](Kimi Team, 2026); [l](Qwen Team, 2026); [m](Dorkenwald et al., 2024). Hosted citations identify the requested public model family. C denotes the connectome transfer probe. Its row uses the same Arena/FAR/FRR/INV equations and a 10,000-draw seed-bootstrap interval under its 1,600-event protocol.

5 Results

5.1 RQ1: Current Evidence Leaves the Policy Target Under-Determined

Table 3 first isolates evidence conditions for three open-weight VLMs, then ranks all 13 models on the shared T+V-8 condition using the response-resolved execution, and finally reports the representative FlyWire task-transfer arm with the same Arena error metrics. Across the ablation, Arena Score stays between 49.1 and 51.4; neither additional frames nor parameter scaling produces a stable gain. The unified leaderboard spans only 49.78–54.01, and all 13 intervals include chance. Most models are strongly deny-biased: allow recall is only 4.9–27.2%, while deny recall is 76.6–95.8%. Resolution recovered all 59 Gemini events, nine of twelve Kimi events, and three of six Grok first-pass failures; the resolved scores and residual INV appear in Table 3 and Appendix C.

Across the three open matrices, no within-model evidence contrast survives Holm correction (adjusted $p \geq .307$): eight frames match one frame, and adding frames to text matches text alone within fold variation.

The Luna–Astra ordering illustrates why score alone is insufficient. Their deny recalls are 85.3% and 85.9%, but Luna recovers 19.5% of resolved allow events versus 14.8% for Astra—16 versus 12 correct allows, with one Astra allow event unresolved. Their FAR is nearly identical (14.7% versus 14.1%), so the 2.04-point Arena gap comes from allow recall. Their intervals overlap broadly. Figure 4 places this recall geometry beside the recurrent-control comparison.

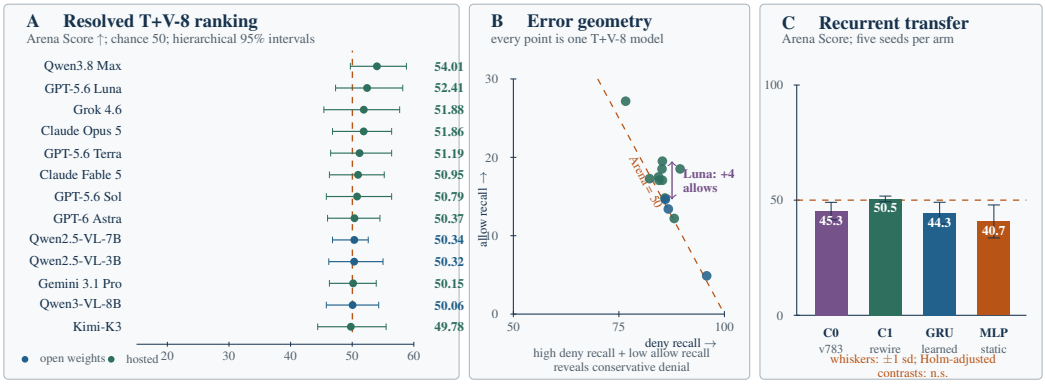

Figure 4: Three complementary views. **A:** response-resolved T+V-8 ranking with hierarchical 95% intervals. **B:** allow versus deny recall exposes the dominant conservative error profile and the Luna–Astra difference. **C:** all four recurrent-control arms are near or below chance, with the degree-preserving graph null highest. Bars show mean Arena Score over five seeds and whiskers show one standard deviation.

5.2 RQ2: Recurrent Capacity Does Not Transfer to Gate Decisions

All 50 synthetic-memory runs are finite and two executions are scientifically identical. The original FlyWire topology (Dorkenwald et al., 2024) reaches capacity 14.828 versus 11.316 for the degree-preserving rewire, a paired difference of 3.512 in 15/15 pairs. At delay 16, their reconstruction scores are 0.882 and 0.002, respectively (Figure 5).

Table 4: Fixed-operator memory results over five sequence seeds. Each null uses three graph seeds; Δ is C0 minus the null. Higher is better.

Condition	Intervention	Test MC [range]	$q(d{=}16)$	Δ [range]
C0	Original v783	14.828 [14.818, 14.838]	0.882	—
C1	Degree-preserving rewire	11.316 [10.971, 11.818]	0.002	3.512 [3.001, 3.856]
C2	Weight shuffle	14.606 [14.416, 14.790]	0.775	0.221 [0.038, 0.403]
C3	Boundary permutation	14.969 [14.961, 14.975]	0.980	−0.141 [−0.153, −0.128]

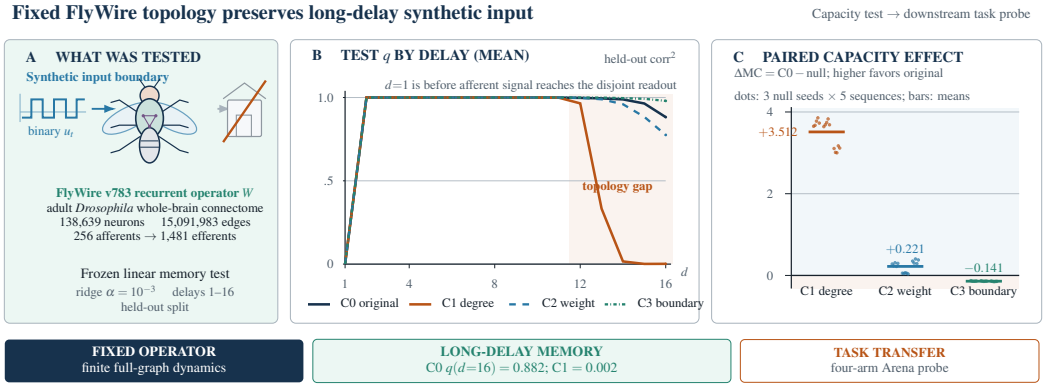

Figure 5: FlyWire v783 retains long-delay synthetic input (B–C), while the task probe stays near chance. Panel A shows the fixed dataflow and a schematic fly icon for the FlyWire operator.

The task-level GPU probe completes four arms by five seeds (20 cells), opening each 1,600-event test split once. No contrast survives Holm correction (Appendix A.2). Under the same decision-metric definitions as the VLM leaderboard, its representative C0 arm obtains Arena 45.25 (seed-bootstrap

95% CI [42.2,48.0]), FAR 74.7%, FRR 34.8%, and INV 0.0% (Table 3, FlyWire row). The original topology carries delayed synthetic input but trails the rewired C1 control by 5.28 Arena points.

Across seeds, Arena Score is 45.25±3.79 for C0, 50.53±1.23 for C1, 44.35±4.73 for the learned GRU, and 40.75±7.16 for the memoryless MLP (mean±sd). Holm-adjusted C0 contrasts are .970, .746, and .094, respectively. Thus neither the biologically grounded topology nor a learned generic recurrent state establishes a downstream advantage.

All cells complete with finite states. The capacity and task probes together separate input retention from task-relevant representation.

5.3 Joint Diagnosis: The Missing Variable Is State, Not More Pixels

The two tracks eliminate different explanations for the same failure. Track P holds events fixed and finds no stable gain from one to eight frames, from vision to text–vision fusion, or from larger model families. Track C shows that a recurrent system can retain delayed input without improving the gate decision. Together, these observations separate three notions that are often conflated: visual coverage, generic memory capacity, and access to the task-defining state.

That distinction matters because the Arena target is not always a function of the current clip. The same visual evidence may be paired with different identity-claim states, while the empty-context leaderboard deliberately withholds the registered and familiarization snapshot. A larger visual window cannot recover a variable that is absent from the observation; recurrence is useful only when it stores the profile and encounter facts on which the policy depends. The benchmark therefore evaluates memory by its lifecycle and contents, not by architectural recurrence alone.

The error decomposition is equally diagnostic. Near-chance Arena Scores coexist with high deny recall because most models reject aggressively; raw accuracy or deny recall alone would overstate capability. Reporting allow recall, FAR/FRR, and class-specific INV alongside the balanced score makes that conservative failure visible and distinguishes policy errors from unresolved outputs.

6 Limitations and Ethical Considerations

The empirical scope is 100 fictional identities, 222 Unreal Engine clips, one Chinese prompt, and one nine-role policy. Synthetic footage controls identity and policy labels but differs from real cameras. Local and hosted stacks use different deterministic frame-extraction implementations, and hosted INV combines output and serving failures. FlyWire results cover one reconstructed topology and one synthetic-input family. The conclusions therefore characterize this corpus, policy, and execution protocol.

7 Conclusion

Gatekeeper Arena defines memory as a lifecycle—profile registration, writable familiarization, then a frozen formal test. In the current-evidence setting, a unified 13-model T+V-8 leaderboard is chance-level and deny-biased. FlyWire retains delayed synthetic input while trailing its rewired control on the gate task. Together, the results identify task-relevant longitudinal state as the central design variable beyond larger visual windows or generic recurrence.

Ethics statement

The corpus uses fictional synthetic characters. Its disclosed policy includes an exclusionary "vagrant" category, and authored visual or linguistic cues may encode stereotypes. Scores measure conformance to that simulation policy; they do not assign moral worth or threat status. The benchmark is scoped to offline research, and release packaging follows the licences governing Unreal Engine and MetaHuman assets.

Reproducibility statement

The supplement provides prompts, fixed identity splits, parsers, metric implementations, bootstrap procedures, and table/figure scripts. Appendix A specifies the scoring and experimental protocol; the release manifest records asset-licence boundaries.

AI use statement

Generative AI tools assisted with research-methodology feedback, experimental design, software implementation and testing, manuscript drafting and editing, literature discovery and summarization, and reference formatting. Inherited project materials report that the synthetic profile narratives were model-authored. Quantitative claims were checked against experimental outputs and automated tests; the authors retain responsibility for the final content.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Agrawal_Dont_Just_Assume_CVPR_2018_paper.html.

Anthropic. Claude Fable 5 and claude Mythos 5. Anthropic Product Announcement, 2026a. URL https://www.anthropic.com/news/claude-fable-5-mythos-5.

Anthropic. Introducing claude Opus 5. Anthropic Product Announcement, 2026b. URL https://www.anthropic.com/news/claude-opus-5.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1563–1572, 2016. doi: 10.1109/CVPR.2016.173. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Bendale_Towards_Open_Set_CVPR_2016_paper.html.

Jiayang Cheng, Dongyu Ru, Lin Qiu, Yiyang Li, Xuezhi Cao, Yangqiu Song, and Xunliang Cai. AMemGym: Interactive memory benchmarking for assistants in long-horizon conversations. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0856bc553d3e3b9827e5140d0ad3bf8d-Abstract-Conference.html.

Joni Dambre, David Verstraeten, Benjamin Schrauwen, and Serge Massar. Information processing capacity of dynamical systems. *Scientific Reports*, 2:514, 2012. doi: 10.1038/srep00514. URL https://doi.org/10.1038/srep00514.

Sven Dorkenwald, Arie Matsliah, Amy R. Sterling, et al. Neuronal wiring diagram of an adult brain. *Nature*, 634(8032):124–138, 2024. doi: 10.1038/s41586-024-07558-y. URL https://doi.org/10.1038/s41586-024-07558-y.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Itai Gat, Idan Schwartz, and Alexander G. Schwing. Perceptual score: What data modalities does your model perceive? In *Advances in Neural Information Processing Systems*, volume 34, pp. 21630–21643, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b51a15f382ac914391a58850ab343b00-Abstract.html.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Google DeepMind. Gemini 3.1 Pro model card. Google DeepMind Model Cards, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Goyal_Making_the_v_CVPR_2017_paper.html.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT, Volume 2 (Short Papers)*, pp. 107–112, 2018. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16077–16102, 2024. doi: 10.18653/v1/2024.acl-long.851. URL https://aclanthology.org/2024.acl-long.851/.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL https://arxiv.org/abs/2607.24653.

Janne K. Lappalainen, Fabian D. Tschopp, Sridhama Prakhya, Mason McGill, Aljoscha Nern, Kazunori Shinomiya, Shin ya Takemura, Eyal Gruntman, Jakob H. Macke, and Srinivas C. Turaga. Connectome-constrained networks predict neural activity across the fly visual system. *Nature*, 634(8036):1132–1140, 2024. doi: 10.1038/s41586-024-07939-3. URL https://doi.org/10.1038/s41586-024-07939-3.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-bench: Benchmarking multimodal LLMs with generative comprehension, 2023. URL https://arxiv.org/abs/2307.16125.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.html.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Zongyu Lin, Zhikun Xu, Xiaohan Song, Yixin Wan, Xingcheng Yao, Tsung-Han Lin, Selina Song, Pranav Subbaraman, Ben Zhou, Kai-Wei Chang, and Yizhou Sun. V-ALPHASOCIAL: Benchmark and self-reflective chain-of-thought generation for visual social commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19025–19047, 2025. doi: 10.18653/v1/2025.findings-acl.975. URL https://aclanthology.org/2025.findings-acl.975/.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. MM-SafetyBench: A benchmark for safety evaluation of multimodal large language models. In *Computer Vision – ECCV 2024*, 2024a. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/7350_ECCV_2024_paper.php.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, pp. 216–233, 2024b. doi: 10.1007/978-3-031-72658-3_13. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/959_ECCV_2024_paper.php.

Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009. doi: 10.1016/j.cosrev.2009.03.005. URL https://doi.org/10.1016/j.cosrev.2009.03.005.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 36, 2023. doi: 10.52202/075280-2004. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets_and_Benchmarks.html.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24868–24891, 2025. doi: 10.18653/v1/2025.emnlp-main.1264. URL https://aclanthology.org/2025.emnlp-main.1264/.

OpenAI. GPT-5.6 system card. OpenAI Deployment Safety Hub, 2026a. URL https://deploymentsafety.openai.com/gpt-5-6.

OpenAI. GPT-6 Astra system card. OpenAI Deployment Safety Hub, 2026b. URL https://deploymentsafety.openai.com/gpt-6-astra/protocolqa-open-ended.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1776–1826, 2022. doi: 10.1145/3531146.3533231. URL https://doi.org/10.1145/3531146.3533231.

Qwen Team. Qwen Code: Model providers. Official Qwen Code documentation, 2026. URL https://github.com/QwenLM/qwen-code/blob/main/docs/users/configuration/model-providers.md.

Philipp Schlegel, Yijie Yin, Alexander Shakeel Bates, et al. Whole-brain annotation and multi-connectome cell typing of Drosophila. *Nature*, 634(8032):139–152, 2024. doi: 10.1038/s41586-024-07686-5. URL https://doi.org/10.1038/s41586-024-07686-5.

Philip K. Shiu, Gabriella R. Sterne, Nico Spiller, et al. A Drosophila computational brain model reveals sensorimotor processing. *Nature*, 634(8032):210–219, 2024. doi: 10.1038/s41586-024-07763-9. URL https://doi.org/10.1038/s41586-024-07763-9.

Sibo Wang-Chen, Victor Alfred Stimpfling, Thomas Ka Chung Lam, Pembe Gizem Özdil, Louise Genoud, Femke Hurtak, and Pavan Ramdya. NeuroMechFly v2: Simulating embodied sensorimotor control in adult Drosophila. *Nature Methods*, 21(12):2353–2362, 2024. doi: 10.1038/s41592-024-02497-y. URL https://doi.org/10.1038/s41592-024-02497-y.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

xAI. Introducing Grok 4.6. xAI Product Announcement, 2026. URL https://x.ai/news/grok-4-6.

Zelai Xu, Zhexuan Xu, Xiangmin Yi, Huining Yuan, Mo Guang, Kaiwen Long, Xinlei Chen, Yi Wu, Chao Yu, and Yu Wang. VS-bench: Evaluating VLMs for strategic abilities in multi-agent environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21561–21572, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Xu_VS-Bench_Evaluating_VLMs_for_Strategic_Abilities_in_Multi-Agent_Environments_CVPR_2026_paper.pdf.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-vet: Evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754, 2024. URL https://proceedings.mlr.press/v235/yu24o.html.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019. doi: 10.1109/CVPR.2019.00901. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zadeh_Social-IQ_A_Question_Answering_Benchmark_for_Artificial_Social_Intelligence_CVPR_2019_paper.html.

Daoan Zhang, Pai Liu, Xiaofei Zhou, Yuan Ge, Guangchen Lan, Jing Bi, Christopher Brinton, Ehsan Hoque, and Jiebo Luo. Visualactbench: Can VLMs see and act like a human?, 2025. URL https://arxiv.org/abs/2512.09907.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, and Jun Zhu. MultiTrust: A comprehensive benchmark towards trustworthy multimodal large language models. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 37, 2024. doi: 10.52202/079017-1561. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/586640cda3db2dc77349013dcefee456-Abstract-Datasets_and_Benchmarks_Track.html.

Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. URL https://arxiv.org/abs/1610.02984.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

A Evaluation Protocol Details

Models, evidence conditions, primary endpoints, and planned contrasts are fixed before scoring. Every condition uses paired encounter schedules. Capability metrics use resolved legal decisions; unresolved scheduled events remain in the coverage denominator and are reported by class. Predeclared contrasts use paired tests with Holm correction. Hosted rows record both requested and returned family aliases.

A.1 Single-Turn Protocol-v1 Contract

The requested response is one JSON object with exactly three fields: `role_guess`, `decision`, and `confidence`. The parser accepts only the nine declared role strings or `unknown`, `allow`/`deny`, and a finite confidence in $[0, 1]$. Parser operations are limited to removing one surrounding Markdown JSON fence; all three fields are then validated directly.

The supplement retains the exact Chinese prompt strings used for every condition and the deterministic message/frame ordering.

A.2 Connectome Memory-Pilot Contract

The memory-capacity pilot uses v783 graph tables from the Shiu et al. companion release and FlyWire neuron annotations v2.1.0. For every delay $d \in \{1, \dots, 16\}$, the target at absolute state index t is u_{t-d}; the washout supplies this history, so no split loses its first 16 rows.

Let Z^{tr} be the efferent states standardized by training means and standard deviations, with every zero-variance feature fixed to zero in all splits, and let Y^{tr} contain the 16 delayed targets. One multi-output readout with intercept is fitted per condition and sequence:

$$(\widehat{B}, \widehat{\beta}) = \arg\min_{B,\beta} \|Y^{\mathrm{tr}} - Z^{\mathrm{tr}} B - \mathbf{1}\beta^{\top}\|_F^2 + 10^{-3}\|B\|_F^2. \tag{6}$$

The ridge coefficient and delay set are fixed before validation. For split $p \in \{\mathrm{val}, \mathrm{test}\}$, condition c, graph or boundary seed g, sequence seed s, and delay d, define

$$q^p_{cgsd} = \mathrm{corr}^2(Y^p_{sd}, \widehat{Y}^p_{cgsd}), \qquad \mathrm{MC}^p_{cgs} = \sum_{d=1}^{16} q^p_{cgsd}. \tag{7}$$

A constant prediction receives $q = 0$; a non-finite state, target, prediction, or score fails the run. Thus each reported memory capacity lies in $[0, 16]$ and retains its complete per-delay decomposition.

C0 uses the original graph and boundaries. C1 performs exact directed degree-preserving rewiring while retaining every source's outgoing signed-weight multiset; C2 retains the edge set but permutes signed weights within source; C3 retains W but jointly reallocates disjoint afferent/efferent labels within full-graph in/out-degree decile strata. C0, C1, and C2 use identical input neurons and signs; C3 applies the same selection rule to its changed afferent pool. Each null seed reuses all five C0 sequence seeds. The frozen paired effect is

$$\Delta^{(k)}_{gs} = \mathrm{MC}^{\mathrm{test}}_{\mathrm{C0},s} - \mathrm{MC}^{\mathrm{test}}_{\mathrm{C}k,g,s}, \quad k \in \{1, 2, 3\}, \tag{8}$$

with C1 the primary topology contrast and C2/C3 explanatory controls. All null families use the complete frozen seed set.

The task-transfer controller uses frozen dimensions I=64, H=16, O=16, giving 3,936 GRU and 1,312 MLP parameters. It consumes the per-event representation under the four arms reported in Section 5.2.

B Additional Scoring and Uncertainty Rules

Every scheduled encounter contributes to coverage. Missing, unparseable, or illegal responses map to $\perp$; capability metrics use resolved decisions and report class-specific coverage.

The encounter is the scoring unit. Define the common event functional

$$Q(g) = \frac{1}{|\mathcal{E}|} \sum_{(i,k)\in\mathcal{E}} \mathbb{1}[g_{ik}], \tag{9}$$

where here $\mathcal{E} = \mathcal{S}_A \cup \mathcal{S}_D$ contains every scheduled encounter. Invalid-output rate is $Q(\neg C)$, where C means that the response satisfies the exact three-field contract. Decision, role, and confidence coverage are respectively $Q(D)$, $Q(R)$, and $Q(H)$, where the named field passes its strict parser. A response may therefore contribute to one field's coverage while failing the exact contract.

Role accuracy applies the same functional to exact agreement where an authored role is available. Macro-F1 is the unweighted mean of the nine role-level F1 values. Invalid or unknown role output counts as a false negative for its authored role.

Track P confidence contributes to field coverage. T+V-8 intervals use 10,000 hierarchical bootstrap replicates: sample folds, then visual-identity clusters within sampled folds.

C Hosted Response Resolution

Every hosted row follows one rule: use a valid first response, then an available valid single-worker recheck; otherwise omit the event from Arena/FAR/FRR and retain it in INV. Rechecks preserve frames, prompt, target, alias, decoding, and parser.

Table 5: Hosted response resolution; residual INV is excluded from capability metrics.

Model	First INV	Rechecked	Recovered	Residual INV	Resolved Arena
Gemini 3.1 Pro Preview	26.22%	59	59	0.00%	50.15
Kimi-K3	5.33%	12	9	1.33%	49.78
Grok 4.6	2.67%	6	3	1.33%	51.88
Claude Fable 5	1.78%	4	0	1.78%	50.95
Claude Opus 5	0.89%	2	0	0.89%	51.86
GPT-6 Astra	0.89%	2	0	0.89%	50.37
GPT-5.6 Sol	0.44%	1	0	0.44%	50.79
Qwen3.8 Max	0.44%	0	0	0.44%	54.01